一种基于知识表示的生物医学实体链接方法

技术领域

本发明属于自然语言处理领域，涉及到一种对生物医学文本进行实体链接的方法，特别涉及到基于知识表示和深度神经网络融合的生物医学实体链接。

背景技术

随着计算机技术和生物技术的高速发展，生物医学领域的文献正在以指数方式增长。面对快速增长的海量数据，研究人员迫切希望揭示蕴含于海量的生物医学文献中的生物医学知识，推动生物医学的发展。这一需求推动了生物医学文本挖掘技术的产生与发展。生物医学命名实体链接（Biomedical Named Entity Linking，BioNEL）作为其中的一项重要研究，目的是促进数据的集成（Data integration）和重用（Re-use）。BioNEL 是指将文本中的生物医学实体（如蛋白质、基因、疾病和药物等）通过知识库映射为唯一标识符（ID），方便于将文本中的非结构化信息转换为结构化数据。它的本质其实是建立文本中实体提及与知识库中实体之间的映射关系，通过建立文本与知识之间的联系，来帮助生物医学知识库自动填充和实体关系抽取等技术的研究。

生物医学命名实体具有如下特点：1）一词多义（多义词），即相同的词或短语可以表示不同的生物命名实体或概念，如：作为生物实体的 CAP 就有多种意义如胱氨酸氨基肽酶(cystine aminopeptidase)、衣壳蛋白(capsid)、环化酶相关蛋白(cyclase-associated protein)和钙激活蛋白(calcium activated protease-Q9UQC9)等；2）多词一义（同义词），即同一生物医学实体具有多种变体。如：PTGS2, cyclooxygenase-2, prostaglandin-endoperoxide synthase 2, COX2 均表示前列腺素过氧化物合成酶。除此之外，生物医学命名实体的缩写被大量使用且不规范，命名方式复杂多样没有统一标准，这都使得生物医学命名实体链接变得

困难。

目前，解决 BioNEL 的方法主要有基于词典的方法，基于向量的方法，基于传统机器学习的方法，以及基于深度学习的方法。

基于词典的方法是通过字符匹配和一些启发式规则从文本中识别词典中的生物医学实体 ID。简单的字符匹配方法可以获得较高的精确率，但是召回率极低。这种情况大多跟上述生物医学命名实体的特点有关。同时，此方法严重依赖词典的完整性和规则的设计，难以被应用于新的领域。

基于向量的方法是将实体提及（Entity Mention）和所有候选实体映射到公共向量空间，然后对每个候选实体定义一个评分度量进行排序（如余弦相似度、欧氏距离、编辑距离、主题相似度、实体流行度等），选取排序第一的候选作为实体提及的链接结果。Leaman 等人（DNorm: disease name normalization with pairwise learning to rank, 2013, Bioinformatics, 29(22): 2909-2917）提出 DNorm 系统，采用向量空间模型来表示医学实体，并使用相似性矩阵来衡量实体提及和候选实体的相似程度。他们在 NCBI 疾病数据集上取得了 0.782 的 F 值，高于基于词典的方法。

基于传统机器学习的方法根据上下文语境对候选实体 ID 进行分类，其目的是对数据的分布进行统计，拟合出数据趋势走向。常用的机器学习模型包括：条件随机域模型(CRF)、支持向量机模型（SVM）、隐马尔可夫模型 (HMM)、最大熵模型(ME)等。但是，基于传统机器学习的方法依赖于复杂的特征工程，需要对数据进行深度探索性分析，根据丰富的领域知识和长期经验来设计和确定模型的最优特征集合，人工成本昂贵且耗时。同时，抽取的特征表示均采用独热（one-hot）的高维稀疏表示方法，难以捕捉文本蕴含的深层语义信息。

基于深度学习的方法克服了对特征工程的依赖，利用多层神经网络构建数

据的深层次抽象特征表示。深度学习的代表性模型主要有自动编码机、RNN、LSTM、CNN 等。Li 等（CNN-based ranking for biomedical entity normalization, 2017, BMC bioinformatics, 18(11):385）将生物医学实体链接任务视为一个排序问题，利用卷积神经网络对候选的语义信息及其形态信息进行建模，然后计算所有<实体，候选>对的相似度得分并排序，得分最高的候选即作为链接结果。他们的模型在 ShARe / CLEF 和 NCBI 数据集上取得了较好的性能。

实体所在的上下文是消歧的关键，正如分布式假说的想法“词的语义由其上下文决定”。上述方法大都着眼于文本数据，通过自动或半自动的方式挖掘有效特征来提高生物医学实体链接的性能。但相比于其它领域，生物医学实体链接需要有力的知识资源支撑，而大量的隐含知识难以在样本数据中进行体现。这些特征数据背后的关联逻辑隐藏在丰富的生物医学词典、知识库（Knowledge Base, KB）等语义网络中，比如蛋白质知识库 UniProt，基因知识库 NCBI Gene 等。他们包含丰富的实体及其结构信息，能够为实体链接任务提供知识支持。然而这些知识在生物医学实体链接系统中尚未得到充分应用。融合实体结构信息和实体语义信息，开展面向大规模生物医学知识库的知识表示学习研究，对于生物医学实体链接具有重要的理论意义和实际应用价值。

发明内容

为利用知识库丰富的实体结构信息帮助克服生物医学实体一词多义和多词一义的难题，本发明提供了一种面向实体结构信息的表示学习方法和一种基于知识表示的生物医学实体链接方法，融合了知识表示与文本语义表示，提高现有生物医学实体链接的性能。

本发明的技术方案：

一种基于知识表示的生物医学实体链接方法，该方法包括三部分：文本预

处理、基于知识库的实体表示学习、基于知识表示的生物医学实体链接。具体步骤如下：

步骤一、文本预处理

对于生物医学文本，首先提取出文本中所有待链接的生物医学实体提及，然后通过知识库查找实体提及对应的所有候选实体标识符（ID）。为了优化内存和运行时间，仅保留排序前五的查找结果作为实体提及的候选 ID 集合。

步骤二、基于生物医学知识库的实体表示学习

知识库中包含了丰富的实体及其结构信息，如同一实体多种变体和不同实体同名。本发明将这些实体结构信息作为向量空间上的约束，采用自动编码机对实体提及表示和变体表示进行重构，从而学习实体 ID 表示。自动编码机是基于如下两个约束：（i）实体 ID 表示是其各个变体表示的和;（ii）实体提及表示是其同名变体表示的和。定义实体提及表示为$m \in \mathbb{R}^d$，变体表示为$v \in \mathbb{R}^d$，实体 ID 表示为$s \in \mathbb{R}^d$。$v^{(i,j)}$是第 i 个实体提及$m^{(i)}$和第 j 个实体 ID$s^{(j)}$共有的变体。则上述自动编码机的两个基本约束可用如下公式表示：

$$s^{(j)} = \sum_i v^{(i,j)}$$

$$m^{(i)} = \sum_j v^{(i,j)}$$

该自动编码机由两部分组成，即编码器和解码器。编码时，编码器按照实体提及→变体→实体 ID 的顺序进行。其中，实体提及表示$m^{(i)}$初始化为其组成单词的预训练词表示的平均值，变体表示$v^{(i,j)}$通过引入一个对角矩阵$E^{(i,j)} \in \mathbb{R}^{d \times d}$对实体提及表示$m^{(i)}$进行分解获得。然后，由对应变体表示的加和获得实体 ID 表示$s^{(j)}$。编码过程的公式如下：

$$s^{(j)} = \sum_i v^{(i,j)} = \sum_i E^{(i,j)} m^{(i)}$$

$E^{(i,j)}$是一个对角矩阵，满足条件$\sum_j E^{(i,j)} = I_n$，其中I_n是一个单位矩阵。

解码时，解码器按照实体 ID→变体→实体提及的顺序进行。通过引入另一个对角矩阵$D^{(j,i)} \in \mathbb{R}^{d\times d}$将编码获得的实体 ID 表示$s^{(j)}$分解为各个变体表示$\bar{v}^{(i,j)}$，然后再由同名变体表示的加和重构实体提及表示$\bar{m}^{(i)}$。解码过程的公式如下：

$$\bar{m}^{(i)} = \sum_j \bar{v}^{(i,j)} = \sum_j D^{(j,i)} s^{(j)}$$

对角矩阵$D^{(j,i)} \in \mathbb{R}^{d\times d}$同样满足条件$\sum_i D^{(j,i)} = I_n$，其中$I_n$是一个单位矩阵。

定义一个重构误差函数来训练自动编码机的参数，其公式为：

$$Loss = \alpha \cdot \| \sum_j (D^{(j,i)} \sum_i E^{(i,j)} m^{(i)}) - m^{(i)} \| + \beta \cdot \| E^{(i,j)} m^{(i)} - D^{(j,i)} s^{(j)} \| \quad \forall i, j$$

该重构误差函数由两部分组成，一个是要求解码出的实体提及表示$\bar{m}^{(i)}$与输入的提及表示$m^{(i)}$对齐，即$\sum_j (D^{(j,i)} \sum_i E^{(i,j)} m^{(i)}) \approx m^{(i)}$；另一个是要求解码器得到的变体表示$\bar{v}^{(i,j)}$与编码器得到的变体表示$v^{(i,j)}$对齐，即$E^{(i,j)} m^{(i)} \approx D^{(j,i)} s^{(j)}$。通过最小化该重构误差函数，使得实体结构信息被嵌入到实体 ID 中，得到学习后的实体 ID 表示。α, β为权重系数且满足$\alpha + \beta = 1$，用于控制两部分对齐的平衡。

综上，生物医学知识库中包含实体间同一实体多种变体和不同实体同名的结构信息，而基于知识库的实体表示学习旨在将这些实体结构信息作为向量空间的约束，通过自动编码机表示为稠密低维实值向量，最终获得实体 ID 表示。

步骤三、基于知识表示的生物医学实体链接

利用步骤二学习获得的实体 ID 表示，对步骤一抽取出的生物医学实体提及进行消歧，获得在特定上下文中实体提及对应的唯一 ID。本发明提供了一种基于知识表示的实体消歧模型，该模型通过注意力机制（Attention）和门机制（Gating）融合文本语义表示和实体 ID 表示，从而预测实体提及被链接到当前候选实体 ID 的概率。具体来说，首先，通过嵌入层，将待链接实体提及的候选实体 ID，和它的左侧上下文、右侧上下文分别映射到向量空间，获得候选 ID 表示s和左、右侧上下文词向量序列$C^L \in \mathbb{R}^{d\times n_2}$和$C^R \in \mathbb{R}^{d\times n_2}$；。然后，$C^L$和$C^R$分别

输入给一个门递归单元神经网络（GRU）获得在第 t 个时间步输出的隐层表示 h_t^L 和 h_t^R 如下：

$$h_t^L = GRU^L\left(C_t^L, h_{t-1}^L\right)$$

$$h_t^R = GRU^R\left(C_t^R, h_{t-1}^R\right)$$

对于一段词序列的语义信息，其中每个词相对于候选 ID 的重要性应该是不同的。为此，我们提出基于知识表示的注意力（Attention）机制，利用候选 ID 表示计算每个时间步隐层表示的归一化权重 α_t。以左上下文表示为例，其计算公式如下：

$$\alpha_t = \frac{\exp\left(e_t\right)}{\sum_k \exp\left(e_k\right)}$$

$$e_t = \tanh\left(W_a^L \cdot h_t^L + V_a^L \cdot s + b_a^L\right)$$

其中，h_t 是 h_t^L 或 h_t^R；$W_a^L \in \mathbb{R}^{1\times d}$，$V_a^L \in \mathbb{R}^{1\times d}$ 和 $b_a^L \in \mathbb{R}^{1\times 1}$ 均是模型的参数，在训练过程中进行调优；s 为当前候选实体 ID 表示；tanh 为双曲正切激活函数。通过一个前馈神经网络建模候选 ID 表示 s 与上下文各个时间步的隐层表示 h_t，获得二者的关联得分 e_t。之后，利用 softmax 函数对得分 e_t 进行归一化得到隐层表示的权重 α_t。

接下来，对 GRU 隐层表示的整个序列作加权求和操作，使结构信息编码的候选 ID 表示与上下文语义表示融合，公式如下：

$$o = \sum_t \alpha_t h_t$$

其中，o 表示左上下文表示 o^L 或右上下文表示 o^R；

注意力机制分别应用于左、右侧隐层表示从而获得的左、右上下文表示 o^L 和 o^R，并通过一个门机制来进行动态控制，让实体提及的最终上下文表示 z 获得充分的学习，计算公式如下：

$$z = g \odot o^L + (1-g) \odot o^R$$

$$g = \sigma\left(W_g \cdot o^L + V_g \cdot o^R + b_g\right)$$

其中，W_g, V_g, b_g 是待训练的参数；⊙表示逐元素相乘；g 是权重，通过将左右上下文 o^L 和 o^R 输入一层全连接层并通过 sigmoid 激活函数 σ 获得的。

最后，我们将实体提及的上下文表示 z 和候选 ID 表示 s 拼接输入给分类器。分类器由具有 ReLU 激活的两层全连接神经网络（FC）和一个包含两个分类单元（分别为被连接和不被链接的概率）的 sigmoid 输出层组成，公式如下：

$$pr_1 = relu\left(W_1 \cdot [z; s] + b_1\right)$$

$$pr_2 = relu\left(W_2 \cdot pr_1 + b_2\right)$$

$$p = sigmoid\left(W_3 \cdot pr_2 + b_3\right)$$

其中，$W_1, b_1, W_2, b_2, W_3, b_3$ 是待训练的参数，$[;]$ 表示拼接操作。

基于知识表示的实体消歧模型通过一个二元交叉熵损失函数进行训练，公式如下：

$$Loss = -\frac{1}{n}\sum_{i=1}^{n}\left(y_i \log p_i + (1-y_i)\log(1-p_i)\right) + \lambda \| W \|$$

其中，n 是训练样例个数，y_i 是第 i 个样例的正确标签，$\lambda \| W \|$ 为训练参数的正则项，p_i 是第 i 个样例的预测概率，。

生物医学实体消歧模型为每个实体提及与其候选 ID 进行打分并排序，选择得分最高的候选 ID 作为最终的链接结果。

本发明的有益效果：

1、本发明利用知识库辅助生物医学实体链接，将知识库中实体间的结构信息（不同实体同名和同一实体多种变体）作为自动编码机的约束，学习实体 ID 表示。该实体 ID 表示嵌入了知识库的实体结构信息，有效解决了实体 ID 表示

质量无法保证的问题，并能同时学习获得多种变体表示和同名实体提及表示。

2、本发明将实体 ID 表示用于生物医学实体链接，构建了一个基于知识表示和深度神经网络结合的生物医学实体链接模型。基于实体 ID 表示，采用注意力机制 GRU 网络，计算待链接实体提及的上下文加权平均表示，深度融合文本语义表示和知识表示两方面的信息，进行实体消歧，有效提高生物医学实体链接的准确性和可靠性。

附图说明

附图 1 是技术流程图。

附图 2 是同一实体多种变体和不同实体同名的语义关系示例图。

附图 3 是自动编码机的结构图。

附图 4 是生物医学实体消歧模型图。

具体实施方式

给定表 1 的一个应用实例，包括文本 1 和文本 2。下面结合本发明的技术方案（和附图）详细叙述本发明的具体实施方式。

表 1 应用实例

文本1	“Dual anti-Ang-2 and anti-**VEGF** therapy acts synergistic on vascular normalization and macrophage infiltration in experimental GBM.”
文本2	“Anti-**VEGF** therapy led to decreased infiltration of CD68+ macrophages in human GBM.”

1、首先，从文本 1 和文本 2 中提取待链接的实体提及“VEGF”。然后通过知识库查找实体提及“VEGF”可能对应的所有候选实体 ID，包括人类属 ID“NCBI Gene: 7422”和家鼠属 ID“NCBI Gene: 22339”。

2、从知识库中抽取同一实体多种变体和不同实体同名等实体结构信息。以图 2 为例，在实线框中，变体“VEGF (human)”，“MVCD1”和“VPF”均表示同一基因（血管内皮生长因子），ID 为“NCBI Gene: 7422”。这就是生物医学实

体的同一实体多种变体问题（同义词）。在虚线框中，变体“VEGF (human)”和“VEGF (mouse)”的名称相同，但是分别对应不同的人类属 ID“NCBI Gene: 7422”和家鼠属 ID“NCBI Gene: 22339”。这就是生物医学实体的不同实体同名问题（多义词）。

3、将这些实体结构信息作为向量空间上的约束，采用自动编码机对实体提及表示和变体表示进行重构，学习基于知识库的实体 ID 表示。自动编码机的结构如图 3 所示，它基于如下两个约束：（i）实体 ID 表示是其各个变体表示的和;（ii）实体提及表示是其同名变体表示的和。以图 2 为例进行说明。根据约束（i），实体 ID“NCBI Gene: 7422”的表示是其各个变体“VEGF (human)”，“MVCD1”和“VPF”表示的和；根据约束（ii），实体提及“VEGF”的表示是其同名变体“VEGF (human)”和“VEGF (mouse)”表示的和。

自动编码机的学习过程如下。首先，按照实体提及→变体→实体 ID 的顺序进行编码。其中，实体提及表示初始化为其组成单词的预训练词向量的平均值；变体表示则通过引入的对角矩阵对同名实体提及表示进行分解获得；实体 ID 表示由其对应的各个变体表示加和获得。然后是解码，按照实体 ID→变体→实体提及的顺序进行。引入另一个对角矩阵，将编码获得的实体 ID 表示分解为各个变体表示，再由同名变体表示的加和重构实体提及表示。自动编码机的目标有两个，一个是让解码部分重构的实体提及表示与输入的提及表示对齐，另一个是让解码部分重构的变体表示与编码部分的变体表示对齐，从而使得他们在向量空间上尽可能接近。最后，通过最小化重构误差函数，调整自编码机的参数，将实体结构信息嵌入到实体 ID 表示中。

3、利用实体消歧模型，通过注意力机制（Attention）和门机制（Gating）融合文本语义表示和实体 ID 表示，预测实体提及被链接到当前候选实体 ID 的

概率。消歧模型如图 4 所示。首先，通过嵌入层，将一个候选实体 ID，和它的左侧上下文、右侧上下文分别映射到向量空间，获得实体 ID 表示和左、右侧上下文词向量序列。然后，将左、右侧上下文词向量序列分别输入给一个门递归单元神经网络（GRU）获得隐层表示。接下来，利用 ID 表示，基于注意力机制计算每个时间步隐层表示的归一化权重，作加权求和获得左、右上下文表示。最后，采用一个门机制组合左、右上下文表示，并与候选 ID 表示拼接输入给分类器。分类器预测实体提及被链接到当前候选实体 ID 的概率，选择概率最高的一个候选 ID 作为最终的链接结果。

在本实例中（表 1），系统识别出文本 1 中的实体提及“VEGF”并将其链接到家鼠属的标识符“NCBI Gene:22339”，识别出文本 2 中的实体提及“VEGF”并将其链接到人类属的标识符“NCBI Gene:7422”上。